RSAC 2025 Conference

Many Voices. One Community.

SESSION ID: ANI-W01

Lessons from AI Red Teaming

…and how to apply them proactively

Daniel Fabian

Digital Arsonist, Google

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

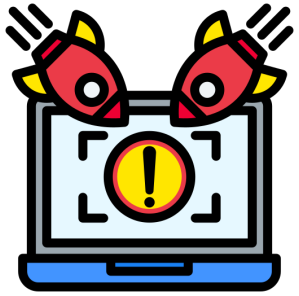
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Overview

AI Red Teaming

The concepts of adversarial testing and adversarial simulation are key in the AI space.



Real-World AI Threats

Threat Intelligence is sparse right now, but attacker motives are predictable.

Focusing on the most realistic attack scenarios helps prioritize.

AI Threat Modelling

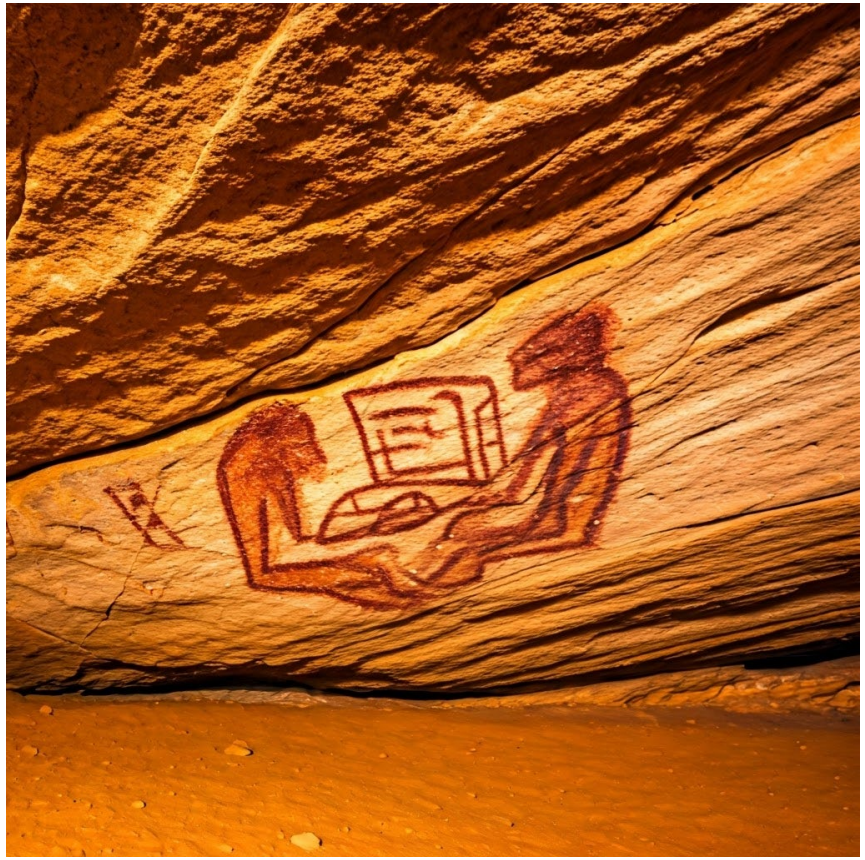
Fixing issues found via Red Teaming is difficult and expensive.

Applying lessons from Red Teaming enables developers to identify pitfalls early.

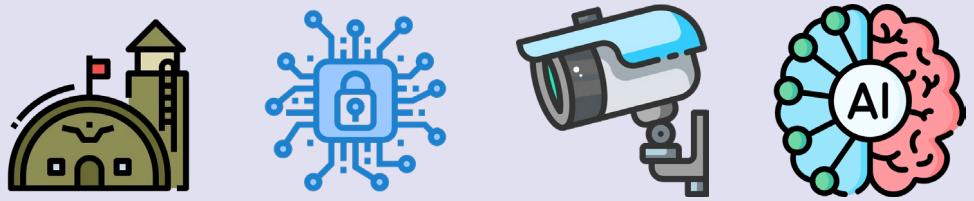
RSAC 2025 Conference

AI Red Teaming

History of Red Teaming



Red Teaming is a **concept** that can be applied to many areas



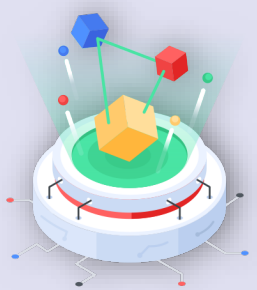
Red Teaming - A Useful Tool

Testing detection & response capabilities

Identifying weaknesses

Raising (security) awareness

What's in the Term "AI Red Teaming"?



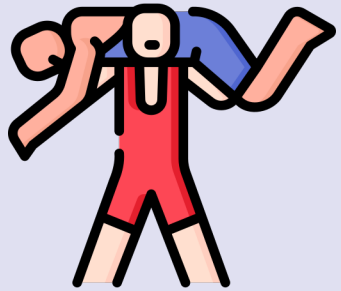
Adversarial Simulation

Executing **end-to-end** attacks

Based on **scenarios**:

- who is the **attacker**?
- what are their **goals**?
- what **capabilities** do they have?

AI specific



Adversarial Testing

Executing **individual** attacks

Typically **narrowly scoped** on specific safety **policy violations**

E.g. prompting for toxicity, bias, and other harms

AI specific

Capabilities Testing

Uncover **harmful/dangerous capabilities**

Areas such as:

- offensive security
- bio, chemical, radiological
- self-replication
- persuasion

Check out the HackTheBox AI Red Teamer Path:
https://academy.hackthebox.com/path/preview/ai-red-teamer

Lessons Learned

Augmenting attacker mindset with adversarial ML expertise

- Starting with classic red team works well
- For more advanced attacks and TTPs, add ML subject matter expertise

Classic security controls help to mitigate findings

- Relying on deterministic security controls to fix issues
- Access control, input/output validation, provenance, etc.

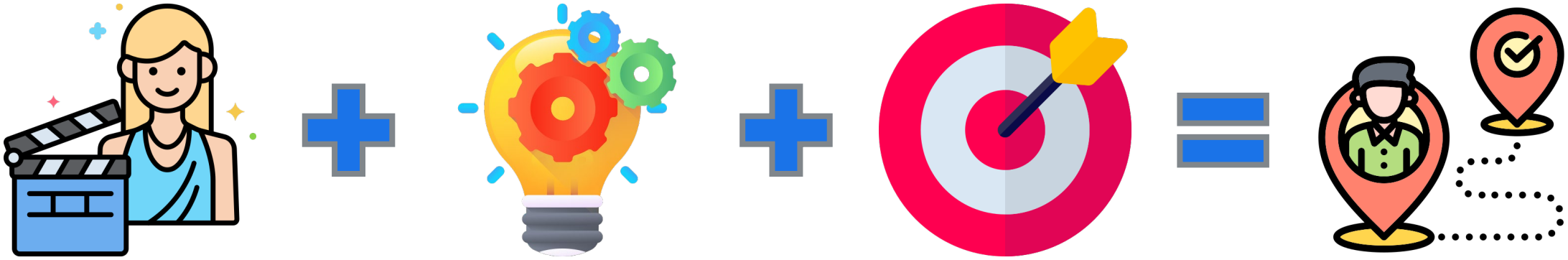
Opportunity to use AI against attackers

- Defensive use of AI capabilities opens up so many possibilities
- Improve both classic security and AI security

RSAC 2025 Conference

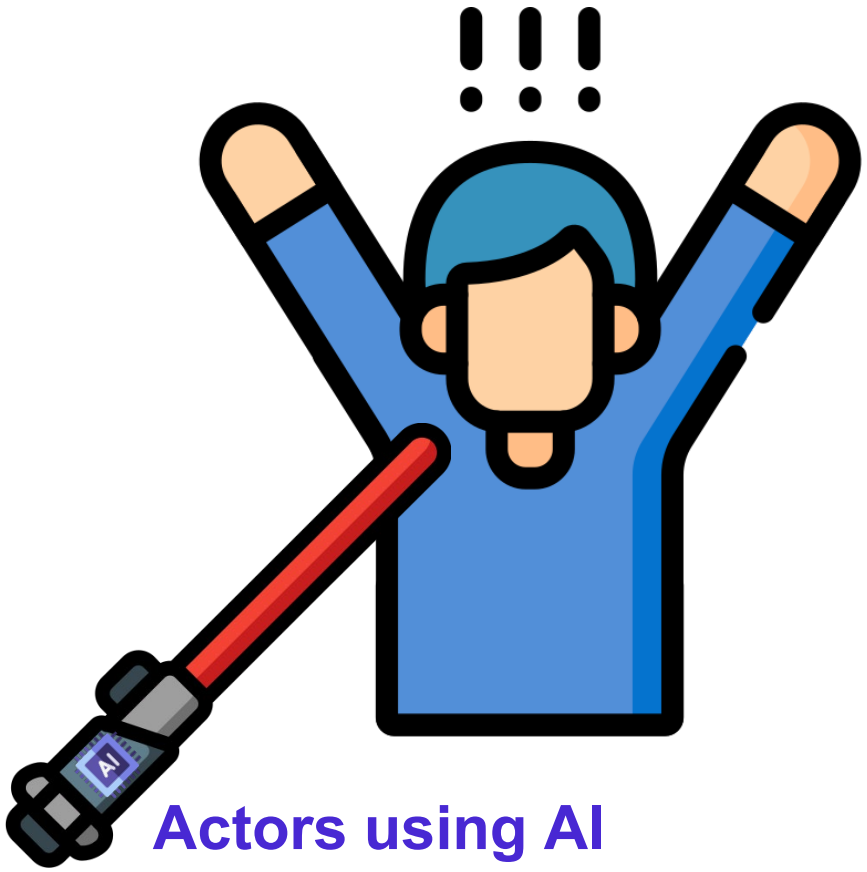
Real-World AI Threats

~~User~~Attacker Journeys

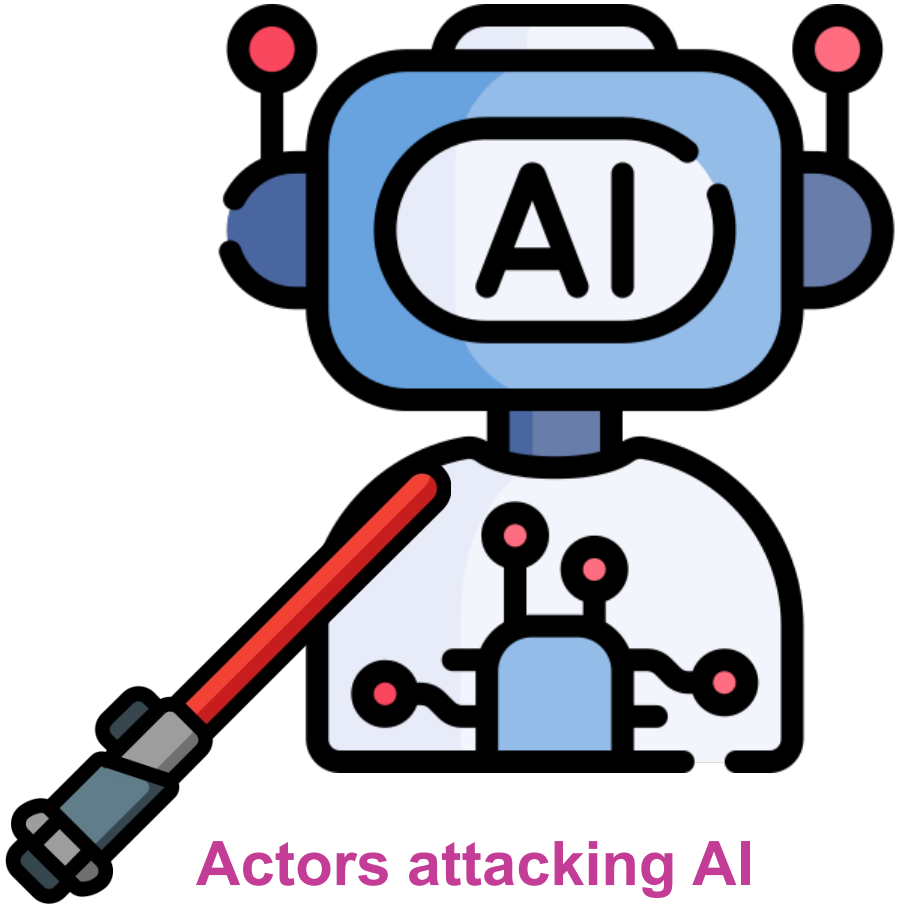


Actor + Motive + Goal = Attacker Journey

Tool vs. Target



Actors using AI for attacks

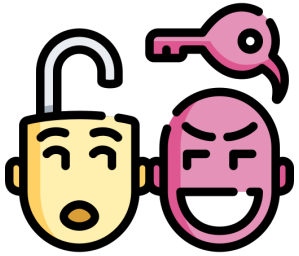


Actors attacking AI

Actors Using AI for Attacks

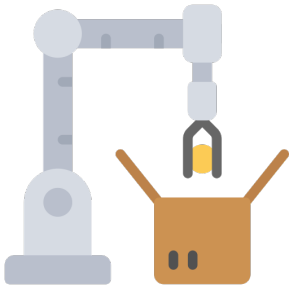
Execution and scaling of information operations

Nation states or activists using AI services to execute and scale information operations related to their cause.



Scaling and automation of social engineering and fraud campaigns

Crime groups and nation states using AI services for scaling and automating social engineering / fraud campaigns.



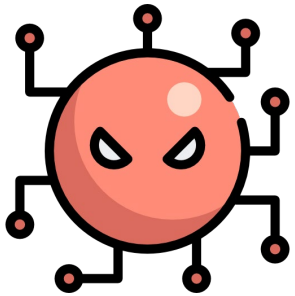
Automation of cyber security attacks

Crime groups developing autonomous agents to find and exploit security vulnerabilities for ransomware / malware distribution.

Actors Using AI for Attacks

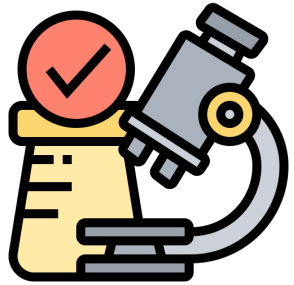
Dissemination of large amounts of stolen information

Actors using AI models to more quickly disseminate large amounts of (stolen) information.



Support for malware development

Actors using code completion and code generation models for malware development.



Cybersecurity research and exploitation

Nation states and crime groups using AI services to identify or develop exploits for new vulnerabilities.

Actors Attacking AI

What do your relevant attacker journeys look like?

[Journey 1]

What can an adversary gain from targeting ***your*** AI capabilities?

[Journey 2]

What (types of) adversaries might be interested in what ***you*** have?

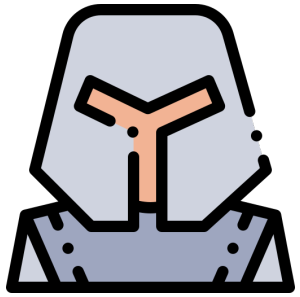
[Journey 3]

What might be the larger motives, and why would **your company** be the easiest way to achieve the adversary's goals?

RSAC 2025 Conference

AI Threat Modeling

Why Do I Want To Do a Threat Model?



It's fun!

Threat modeling gives you an **opportunity to cosplay as an attacker**. You have permission to think about what you would do if you were the attacker.

It protects the company and its customers

Being aware of what could go wrong is the first step for **launching a product that is safe** for users and protects the company's reputation.

It avoids nasty surprises later on

Helps identify risky areas and potential pitfalls **early in the development lifecycle**, when it's still cheap to fix issues.

Broad Categories of AI Risks

Content safety risks

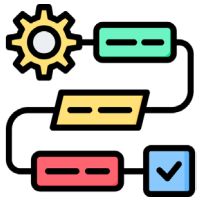
Could users be exposed to unsafe content in your application?

Infrastructure risks

How might an attacker target the infrastructure used along the model lifecycle

Integration risks

Is the model safely integrated into your application, avoiding common issues?



Process risks

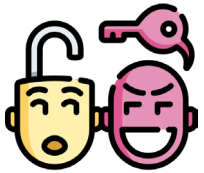
What can go wrong in development and maintenance processes?

Content Safety Risks

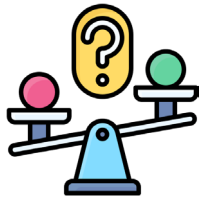
Hallucinations

Dangerous capabilities

Jailbreaks



Social engineering



Fairness issues or bias

Integration Risks

Indirect prompt injection
goo.gle/camel

Control sequence injection

Privilege abuse

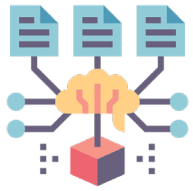
Memory smuggling / theft

Malicious tool use

Prompt spilling / leakage

Adversarial examples / evasion risks

Infrastructure Risks



Model theft

Training data theft

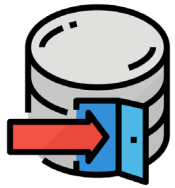
Training data leakage

Process Risks

Unauthorized training data

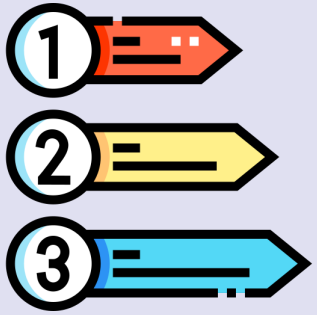
(Lack of) model patching

Training data poisoning



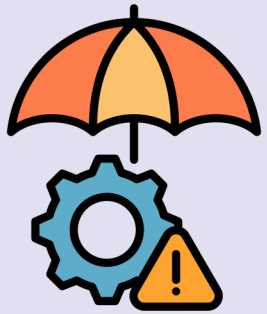
Model backdoors

So Now I Have a Threat Model…



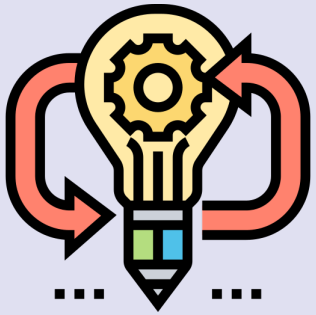
Prioritize

Not all threats are created equal. Ranking threats based on likelihood and potential impact helps **focus mitigation efforts**.



Mitigate

Think about what controls you can implement to **reduce the risk** coming from the main identified threats.



Iterate

Regularly review and update as your AI application evolves and new threats emerge.

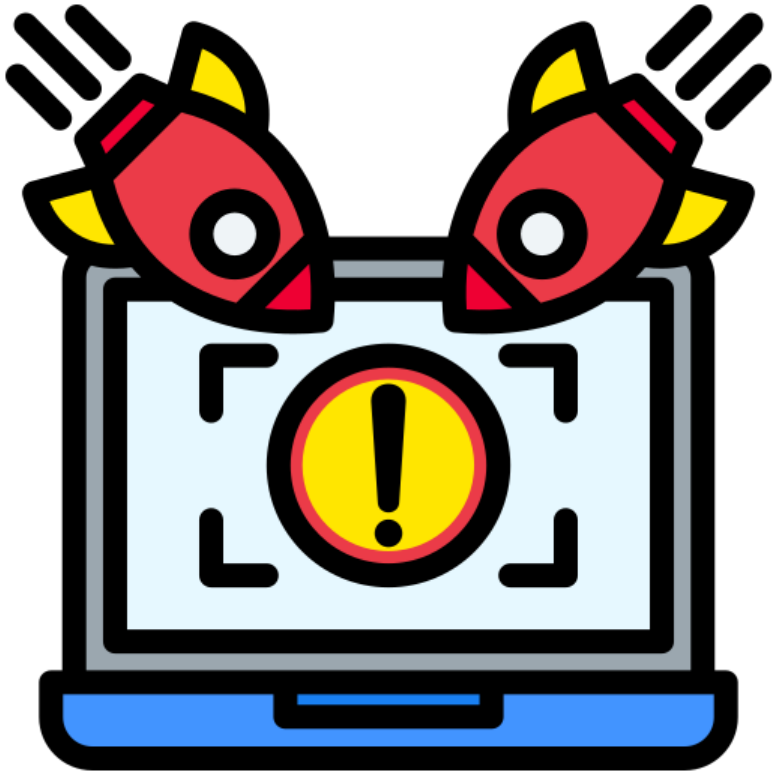
RSAC 2025 Conference

Key Takeaways / Apply

AI Red Teaming

- Think about:
  - What are the things I want to achieve through Red Teaming?
  - Is Red Teaming the most appropriate tool for this purpose?
  - What types of AI Red Teaming does my organization need?

Real-World AI Threats



- Think about:
  - What do I have to "offer" to potential attackers?
  - What are the key attacker journeys I need to worry about?
  - How can these attacker journeys help me prioritize security and safety efforts?

AI Threat Modelling

- Think about:
  - Have I considered the most important threats in the AI space and how they can affect my application?
  - Are all my engineers aware too?
  - How can I integrate threat modelling into my SDLC so I can catch issues early?
  - Do I regularly update and augment my threat model?

RSAC 2025 Conference

Q&A

RSAC™ | 2025 Conference

Many Voices.
One Community.